

# Trajectory Conditioned Cross-embodiment Skill Transfer

Yuhang Tang<sup>1,2</sup> Yixuan Lou<sup>1</sup> Pengfei Han<sup>2</sup> Haoming Song<sup>2,3</sup>

Xinyi Ye<sup>2</sup> Dong Wang<sup>2</sup> Bin Zhao<sup>†1</sup>

<sup>1</sup>Northwestern Polytechnical University <sup>2</sup>Shanghai AI Laboratory <sup>3</sup>Shanghai Jiao Tong University

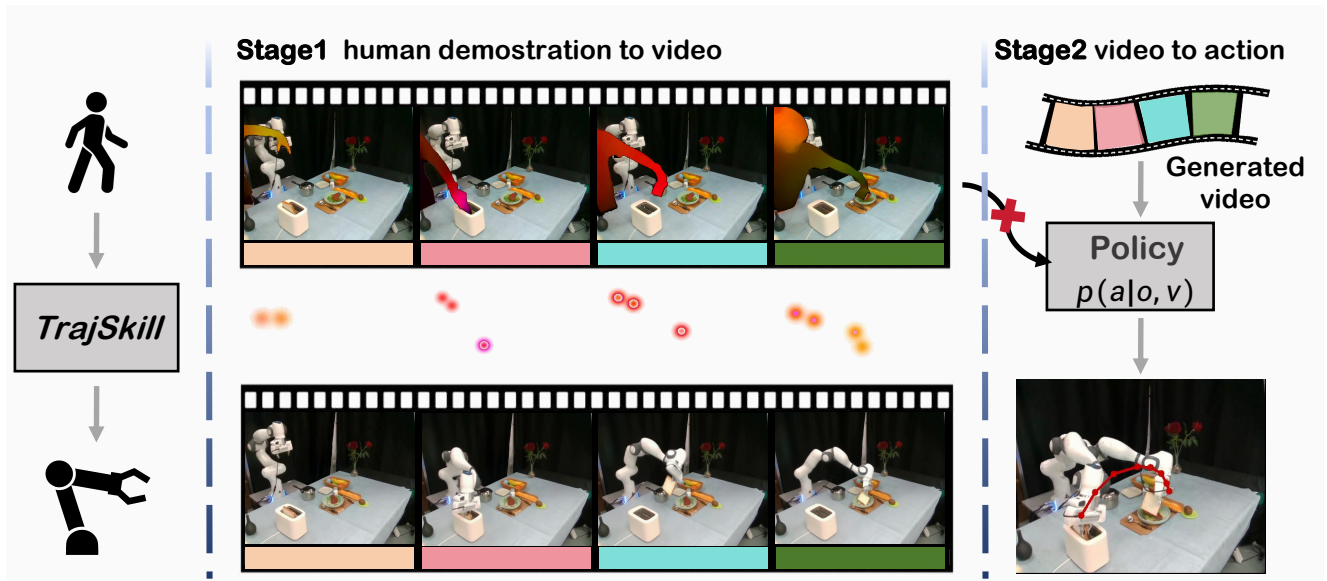


Fig. 1: Overview of the TrajSkill from human to robot action. TrajSkill leverages sparse optical flow as a universal motion representation, achieving zero-shot imitation without reinforcement learning or paired datasets. In Stage 1, dense optical flow is extracted from human demonstrations and sampled into sparse optical flow to guide video generation. In Stage 2, the generated video is translated into robot actions using a learned policy, enabling the robot to mimic the demonstrated task.

**Abstract**—Learning manipulation skills from human demonstration videos presents a promising yet challenging problem, primarily due to the significant embodiment gap between human body and robot manipulators. Existing methods rely on paired datasets or hand-crafted rewards, which limit scalability and generalization. We propose TrajSkill, a framework for *Trajectory Conditioned Cross-embodiment Skill Transfer*, enabling robots to acquire manipulation skills directly from human demonstration videos. Our key insight is to represent human motions as *sparse optical flow trajectories*, which serve as embodiment-agnostic motion cues by removing morphological variations while preserving essential dynamics. Conditioned on these trajectories together with visual and textual inputs, TrajSkill jointly synthesizes temporally consistent robot manipulation videos and translates them into executable actions, thereby achieving cross-embodiment skill transfer. Extensive experiments are conducted, and the results on simulation data (MetaWorld) show that TrajSkill reduces FVD by 39.6% and KVD by 36.6% compared with the state-of-the-art, and improves cross-embodiment success rate by up to 16.7%. Real-robot experiments in kitchen manipulation tasks further validate the effectiveness of our approach, demonstrating practical human-to-robot skill transfer across embodiments.

## I. INTRODUCTION

Robotic manipulation learning from human demonstration videos has been a long, compelling yet challenging task

in embodied intelligence. While Human videos naturally capture manipulation dynamics, the direct transfer of these skills remains impracticable due to substantial differences in morphology, kinematic constraints, and embodiment between the human body and robot manipulators. Previous approaches have attempted to bridge this gap through reinforcement learning with hand-crafted reward functions [1], [2], meta-learning for one-shot imitation [3], or domain alignment techniques between human and robot embodiments [4]. However, these methods often depend on costly human interventions, paired datasets, or brittle alignment strategies, which limits their scalability and practical deployment in real-world scenarios.

Recent advances in video generation models open new avenues for robot policy learning [5], offering the potential to synthesize long-horizon motion sequences that can inform planning [6], [7]. Parallel to these developments, a growing body of work has sought to directly learn robot policies from human demonstration videos. For example, Learning by Watching [8] translates human motions into robot actions via keypoint-based representations, but relies on accurate mappings between embodiments. Vid2Robot [9] introduces an end-to-end video-conditioned policy that learns from paired

human videos and robot trajectories, yet its dependence on large paired datasets and embodiment alignment remains a bottleneck. More recently, Human2Robot [10] formulates video-to-action transfer as a diffusion-based generative task on a large paired dataset, while Motion Tracks [11] introduces a keypoint retargeting network for few-shot transfer; both approaches, however, remain sensitive to embodiment discrepancies. Despite these advances, existing methods typically require paired datasets, hand-crafted rewards, or explicit human-robot alignment strategies, all of which stem from the fundamental embodiment gap between human and robot morphologies. This reliance limits their scalability and hinders the development of morphology-invariant representations for robust skill transfer.

To solve aforementioned challenges, it is essential to discover a motion representation that is compact, embodiment-agnostic, and retains essential task dynamics. We observe that employing sparse optical flow effectively filters out appearance and morphological differences while preserving the key motion intent, thereby achieving embodiment invariance. Building on this insight, we propose **TrajSkill**, a trajectory conditioned cross-embodiment skill transfer framework. TrajSkill leverages *sparse optical flow trajectories* extracted from human demonstrations as a unified motion representation, which eliminates embodiment-specific appearance while preserving dynamic motion patterns. Conditioned on these trajectories, TrajSkill converts human demonstrations into executable robot policies, achieving zero-shot imitation without reinforcement learning or paired datasets.

Our contributions are summarized as follows:

- We propose TrajSkill, a framework for trajectory conditioned cross-embodiment skill transfer, which jointly enables controllable video generation and executable robot policy learning directly from human demonstration videos.
- We introduce sparse optical flow trajectories as an embodiment-agnostic representation that bridges the morphological gap between human and robot embodiments, providing effective motion cues for skill transfer.
- We validate TrajSkill extensively across a diverse set of manipulation benchmarks encompassing dozens of tasks, demonstrating consistent improvements in video generation quality, cross-embodiment success rates, and real-robot skill execution in challenging kitchen manipulation scenarios.

## II. RELATED WORK

### A. Video Diffusion Models

Video Diffusion Models (VDMs) have recently achieved impressive progress in generating high-quality video content. Early methods extended image diffusion architectures by adding temporal convolutions and attention layers within a UNet backbone [12], [13]. While these approaches demonstrated the feasibility of diffusion-based video synthesis, their scalability and long-horizon consistency were fundamentally constrained. Subsequent works such as VideoCrafter [14]

and Stable Video Diffusion [15] expanded training to larger datasets but still struggled with generating temporally coherent long sequences.

The introduction of Diffusion Transformers (DiT) marked a paradigm shift, enabling more scalable and unified sequence modeling [16]. Large-scale systems such as Sora [6], Vidu [5], and CogVideoX [7] demonstrate the capability of DiT to generate high-definition videos extending to tens of seconds or more, with flexible aspect ratios and improved motion fidelity. Building upon these advances, our work adopts a DiT backbone for trajectory conditioned video synthesis with enhanced temporal coherence.

### B. Motion Control in Video Generation and Robotic Manipulation

Beyond generating realistic appearance, controllable motion generation is critical for both video synthesis and robotics. In video generation, prior works have proposed conditioning on reference videos [17], [18], structural cues such as depth maps or sketches [19], or object masks [20], [21]. Recently, trajectory-based conditioning has attracted attention due to its physical intuitiveness, allowing users to directly specify object or camera motion [22], [23]. However, these methods often struggle with motion consistency over long horizons.

In robotics, generative video models have been explored as policy representations, where predicted video plans are mapped into executable actions. Works such as UniSim [24] and UniPi [25] use text- or image-conditioned video prediction for robot interaction planning, while SuSIE [26] and AVDC [27] incorporate autoregressive or custom diffusion architectures to infer actions from predicted trajectories or optical flow. More recent approaches such as SEER [28] and This&That [29] leverage language and multimodal signals for temporally aligned video generation. These works underscore the potential of video diffusion for scalable policy learning, yet they leave unresolved the challenge of generating robot-consistent motion videos with high fidelity and controllable trajectories.

### C. Cross-embodiment Learning from Human Videos

Learning from human demonstrations offers a scalable alternative to costly robot-collected data. Prior methods construct rewards from human videos [1], [2], perform one-shot imitation via meta-learning [3], or learn aligned visual embeddings across human and robot domains [30]. Others extract affordance cues [31], human-object interactions [32], or explicit hand trajectories and keypoints [33], [34]. Despite recent progress, these approaches often rely on reinforcement learning loops, require paired datasets, or suffer from brittle trajectory retargeting due to morphological mismatches.

To address these challenges, we introduce sparse optical flow trajectories as an embodiment-agnostic motion representation. By projecting both human and robot motions into a unified 2D trajectory space and conditioning a Diffusion Transformer-based video generator on these signals, we

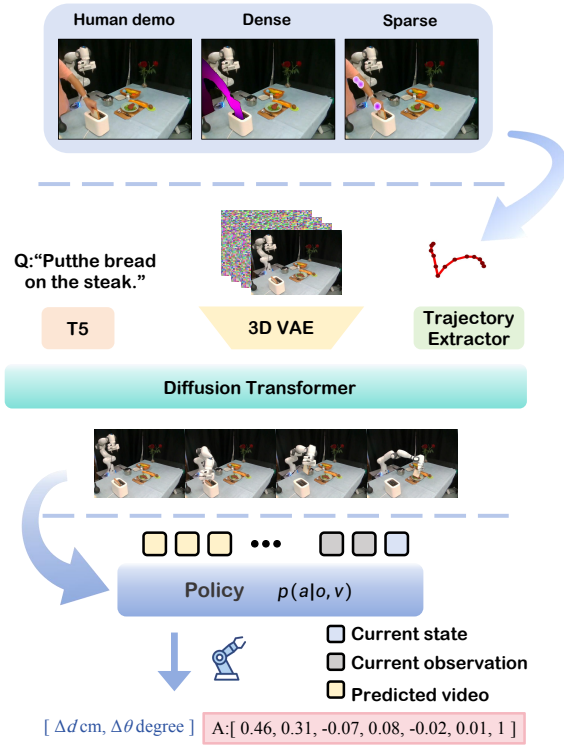


Fig. 2: Unified illustration of the TrajSkill framework. **Top:** Embodiment-Invariant Flow Sampling. From a human demonstration video frame (left), dense optical flow is computed by RAFT [35] (middle), and sparse keypoint trajectories are sampled according to the flow magnitude and propagated over time (right). **Middle and Bottom:** Overview of the Trajectory Conditioned Robot Execution. Given a task description, the T5 model interprets the instruction, a 3D VAE extracts spatial features, and the trajectory extractor provides sparse flow signals. These are fused within a Diffusion Transformer to predict robot motion videos, which are then decoded by the policy  $p(a|o, v)$  into executable actions.

achieve cross-embodiment skill transfer without reinforcement learning optimization or paired training data, enabling scalable one-shot imitation from human demonstrations.

### III. METHOD

Our framework realizes trajectory conditioned cross-embodiment skill transfer by leveraging sparse optical flow as an embodiment-agnostic representation. TrajSkill consists of three components: (1) embodiment-invariant flow sampling, (2) trajectory conditioned robot execution, and (3) cross-embodiment skill transfer.

#### A. Embodiment-Invariant Flow Sampling

As shown in the top of Fig. 2, our Embodiment-Invariant Flow Sampling computes dense optical flow and reduces it to compact sparse trajectories. Given a human demonstration video, we first compute the dense optical flow  $\mathbf{F}_t \in \mathbb{R}^{H \times W \times 2}$  between consecutive frames using RAFT [35]. The flow  $\mathbf{F}_t$  represents the displacement between pixel locations across

frames. To construct a compact trajectory representation, we define a grid of candidate positions with a stride  $\lambda$ :

$$C = \{(x, y) \mid x \in \{o_w, o_w + \lambda, \dots, W\}, y \in \{o_h, o_h + \lambda, \dots, H\}\}, \quad (1)$$

where  $o_w$  and  $o_h$  are random offsets within the image dimensions  $H \times W$ . Each candidate is sampled with probability proportional to its initial flow magnitude. Specifically, the probability for a candidate  $(x, y)$  is:

$$p_{(x,y)} = \frac{\|\mathbf{F}_0(x, y)\|_2}{\sum_{(x', y') \in C} \|\mathbf{F}_0(x', y')\|_2}, \quad (2)$$

where  $\|\mathbf{F}_0(x, y)\|_2$  is the  $L_2$ -norm of the flow vector at position  $(x, y)$  in the first frame. The probabilistic sampling ensures that regions with stronger motion are more likely to be selected as candidate positions, effectively focusing on areas with significant movement.

To determine the actual sampled keypoints, we first draw the number of samples  $N$  uniformly from a maximum budget  $N_{\max}$ :

$$N \sim \text{Uniform}\{1, \dots, N_{\max}\}, \quad (3)$$

and then select  $N$  distinct candidates without replacement according to the probability distribution  $p_{(x,y)}$ . The initial keypoint set:

$$K_0 = \{(x_{u_1}, y_{u_1}), \dots, (x_{u_N}, y_{u_N})\}, \quad (x_{u_k}, y_{u_k}) \sim p_{(x,y)}. \quad (4)$$

Next, we propagate the selected keypoints through time by integrating the local flow vectors. At each timestep  $t$ , the new position  $(x^{t+1}, y^{t+1})$  of a keypoint is updated using the flow vector at current position  $(x^t, y^t)$ :

$$(x^{t+1}, y^{t+1}) = (x^t, y^t) + \mathbf{F}_t(x^t, y^t), \quad (5)$$

where  $\mathbf{F}_t(x^t, y^t)$  is the flow vector at position  $(x^t, y^t)$  at time  $t$ . By repeating this process, we obtain a set of sparse trajectories  $T$ :

$$T = \{(x'_i, y'_i)\}_{i,t}, \quad (6)$$

where each trajectory  $T_i = \{(x'_i, y'_i)\}_t$  represents the path of a selected keypoint over time.

Finally, to mitigate noise and improve spatial consistency, we apply a Gaussian blur smoothing to the sparse flow field before usage. Concretely, each flow channel is convolved with a normalized isotropic Gaussian kernel:

$$\tilde{S}_t^{(d)}(u, v) = \sum_{i=-k}^k \sum_{j=-k}^k G(i, j) S_t^{(d)}(u-i, v-j), \quad (7)$$

where  $S_t^{(d)}$  denotes the sparse flow channel  $d \in \{x, y\}$ ,  $G(i, j)$  is the discrete Gaussian kernel. The smoothed flow  $\tilde{S}_t = (\tilde{S}_t^{(x)}, \tilde{S}_t^{(y)})$  is then used to construct the final trajectory representation. It discards the embodiment-specific appearance details, focusing solely on the essential motion intent, which is key for analyzing the human demonstration while maintaining compactness and efficiency in trajectory representation. In this way, the embodiment gap is eliminated, ensuring that the representation consistently reflects task dynamics rather than morphological variations.

## B. Trajectory Conditioned Robot Execution

**Trajectory Conditioned Video Generation** To synthesize robot manipulation sequences conditioned on instruction and trajectory inputs, we adopt a latent diffusion transformer (DiT) backbone [7]. Unlike prior UNet-based video diffusion models that rely on local convolutional receptive fields, our architecture leverages global attention to capture long-range temporal dependencies, thereby enabling scalable modeling of long-horizon robot motion with improved temporal coherence. As shown in the middle of Fig. 2, the process is defined as:

$$V_r = G(I_0, C_{\text{text}}, C_{\text{traj}}), \quad (8)$$

where  $I_0$  denotes the initial frame,  $C_{\text{text}}$  represents the task instruction, and  $C_{\text{traj}}$  provides sparse trajectory signals.

To bridge the gap between dense motion learning and sparse trajectory conditioning, we introduce a two-stage training strategy:

- **Stage 1: Dense Flow Supervision.** The model is first trained with dense optical flow, providing detailed motion cues that enable learning of accurate robot dynamics and object interactions.
- **Stage 2: Sparse Trajectory Alignment.** Training the transitions to sparse flow trajectories, aligning supervision with inference conditions and ensuring morphology-invariant motion control.

The two-stage design allows the generator to first acquire precise motion priors and subsequently adapt to sparse trajectory prompts, ensuring that generated videos not only follow human-demonstrated intent but also generalize across embodiments.

**Video Policy to Robot Execution** The generated video  $V_r$  must be mapped to executable robot actions. As illustrated in the bottom of Fig. 2, the policy is conditioned on both the current observation and the predicted video. To incorporate  $V_r$ , the video frames are temporally aggregated into a compact reference image, which is subsequently projected into the model space and fused with the current state embedding. The fused representation is then decoded into a action sequence:

$$A_r = F(O, S, V_r), \quad (9)$$

where  $O$  denotes the robot’s real-time observation,  $S$  the low-level state information,  $V_r$  the generated video, and  $F(\cdot)$  the policy network integrating all inputs.

## C. Cross-embodiment Skill Transfer

The preceding components together establish the foundation for cross-embodiment transfer. First, the embodiment-invariant flow sampling module extracts sparse optical flow trajectories that abstract away embodiment-specific appearance and kinematics. Second, the two-stage training pipeline realizes trajectory conditioned robot execution, where sparse trajectories from human demonstrations guide the generation of robot-executable task videos, which are subsequently translated into actions through video-policy to robot execution. By combining these two components, our framework

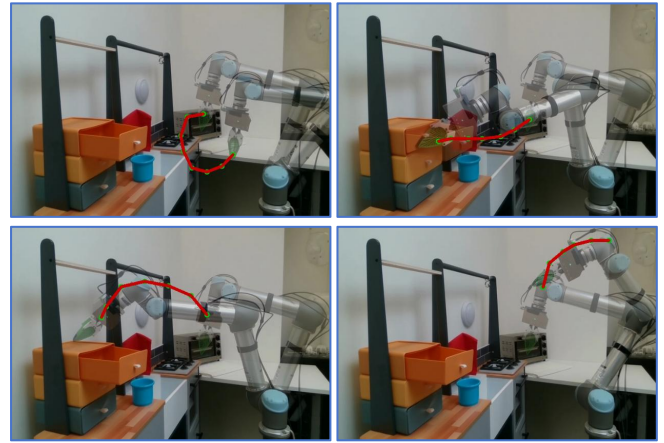


Fig. 3: Trajectory conditioned video generation. The robot arm is provided with an initial frame and a predefined trajectory, shown as red curves. TrajSkill generates a sequence of motion frames where the robot follows the specified path. The figure illustrates the robotic arm at both the starting and ending points of the trajectory.

enables one-shot cross-embodiment imitation. The formulation allows a robot to reproduce human-demonstrated skills without paired datasets or reinforcement learning, bridging the embodiment gap and enabling scalable trajectory conditioned skill transfer.

## IV. EXPERIMENTS

We design comprehensive experiments to evaluate the proposed TrajSkill framework in terms of (i) trajectory controllability in generated videos, (ii) cross-embodiment transfer from human demonstrations to robots, and (iii) robot execution for downstream manipulation tasks.

### A. Experimental Setup

1) *Datasets*: We conduct evaluations on three representative benchmarks:

- **MetaWorld 50 Tasks** [36]: a suite of 50 simulated manipulation tasks on a Sawyer arm, covering four difficulty levels (*easy*, *medium*, *hard*, *very hard*) [37].
- **Franka Multi-Tasks** [38]: real-world demonstrations across 14 diverse tasks with a Franka Panda robot.
- **XSkill** [39]: a hybrid dataset supporting cross-embodiment transfer, containing both sphere-agent trajectories in simulation and real human-hand demonstrations.

2) *Baselines*: We benchmark against both video generation and robot execution approaches:

- **Video Generation**: AVDC [27], This&That [29], and CogVideo [7].
- **Robot Execution**: Diffusion Policy (DP) [40], TinyVLA [41], SmolVLA [42], and OCTO [43].

3) *Evaluation Metrics*: We evaluate the generated videos using two key metrics: Fréchet Video Distance (FVD) [44] and Kernel Video Distance (KVD) [28], which measure the realism and temporal consistency of the videos. Additionally,

Method	Publication	MetaWorld		Franka		Resolution	Frames
		FVD ( $\downarrow$ )	KVD ( $\downarrow$ )	FVD ( $\downarrow$ )	KVD ( $\downarrow$ )		
AVDC (V+Lang.)	<i>ICLR 2024</i>	1467.68	1632.93	925.27	775.19	128×128	8
This&That (V+Lang.)	<i>ICRA 2025</i>	857.08	796.31	991.87	987.20	448×448	28
CogVideo (V+Lang.)	<i>ICLR 2025</i>	528.04	422.86	427.30	255.82	480×720	49
Ours (V+Lang.+Traj)	—	<b>318.83</b>	<b>268.03</b>	<b>309.53</b>	<b>175.01</b>	480×720	49

TABLE I: Quantitative evaluation on MetaWorld and Franka. Results are reported in terms of FVD and KVD (lower values indicate better performance). TrajSkill consistently outperforms all baselines.

we assess the Success Rate (SR) [45], which represents the percentage of completed manipulation tasks and serves as the ultimate metric for skill transfer.

### B. Trajectory Conditioned Video Generation

1) *Visual Quality*: We evaluate on 57 test sequences sampled from MetaWorld and Franka datasets. Table I shows that TrajSkill achieves the best FVD and KVD across datasets, outperforming both robotics-specific VDMs and large-scale video generators. TrajSkill reduces FVD by 39.6% and KVD by 36.6% on MetaWorld, and by 27.6% (FVD) and 31.6% (KVD) on Franka compared to CogVideo. Moreover, our method produces longer and clearer videos, crucial for manipulation planning.

2) *Trajectory Controllability*: To further evaluate the controllability of our framework, we condition the video generator on trajectories and measure whether the generated robot motions remain faithful to the given paths. As illustrated in Fig. 3, the generated motions closely align with the input trajectories across both simple linear paths and more complex curved patterns.

From a cross-embodiment perspective, this result highlights that the 2D trajectory abstract away embodiment-specific morphology yet still convey precise spatiotemporal guidance. In practice, this means that demonstrations performed by human hands can be faithfully reinterpreted into robot-consistent motion videos, effectively bridging the gap between human and robotic embodiments.

### C. Cross-embodiment Skill Transfer

1) *Simulation Transfer (Sphere  $\rightarrow$  Robot)*: Using XSkill’s sphere-agent demonstrations, we extract sparse flow trajectories to condition robot motion generation. As shown in Fig. 4 (top), generated videos follow the demonstrated motions, confirming accurate simulation-to-robot transfer.

2) *Real-World Transfer (Human  $\rightarrow$  Robot)*: We further evaluate transfer from human hand demonstrations. Sparse trajectories extracted from human videos effectively guide robot motion generation, as shown in the middle and bottom of Fig. 4. These results demonstrate that TrajSkill bridges embodiment gaps in both simulated and real-world scenarios.

Method	Success Rate (%)				
	Easy	Medium	Hard	Very Hard	Overall
Diffusion Policy	23.1	10.7	1.9	6.1	10.5
TinyVLA	77.6	21.5	11.4	15.8	31.6
SmolVLA	74.6	<b>30.9</b>	18.0	30.0	38.3
Ours	<b>81.8</b>	29.1	<b>38.0</b>	<b>30.0</b>	<b>44.7</b>

TABLE II: Success rate comparison across task difficulties. Performance comparison of different methods on 49 robot tasks categorized by difficulty levels, showing success rates in percentage.

Method	Octo	Diffusion Policy	Ours
Pick	0.0%	81.8%	<b>90.9%</b>
Place	0.0%	72.7%	<b>81.8%</b>

TABLE III: Real robot experimental results on ”Put the Banana in the Basket” task. Success rates for Pick and Place actions across different methods.

These cross-embodiment experiments highlight TrajSkill’s ability to generalize motion knowledge across distinct agents, from simulated spheres to robots and from human hands to robotic manipulators. By demonstrating consistent trajectory transfer in both controlled and real-world settings, we establish a strong foundation for translating generated video policies into executable robot actions. Building on this, the following section investigates how such transferred skills materialize in actual robot execution, validating their practicality and robustness in complex environments.

### D. Robot Execution

1) *Simulation Rollouts*: We evaluate robot execution of TrajSkill on the MetaWorld 50 benchmark. As reported in Table II, TrajSkill achieves the highest overall success rate (44.7%), outperforming prior approaches by a clear margin.

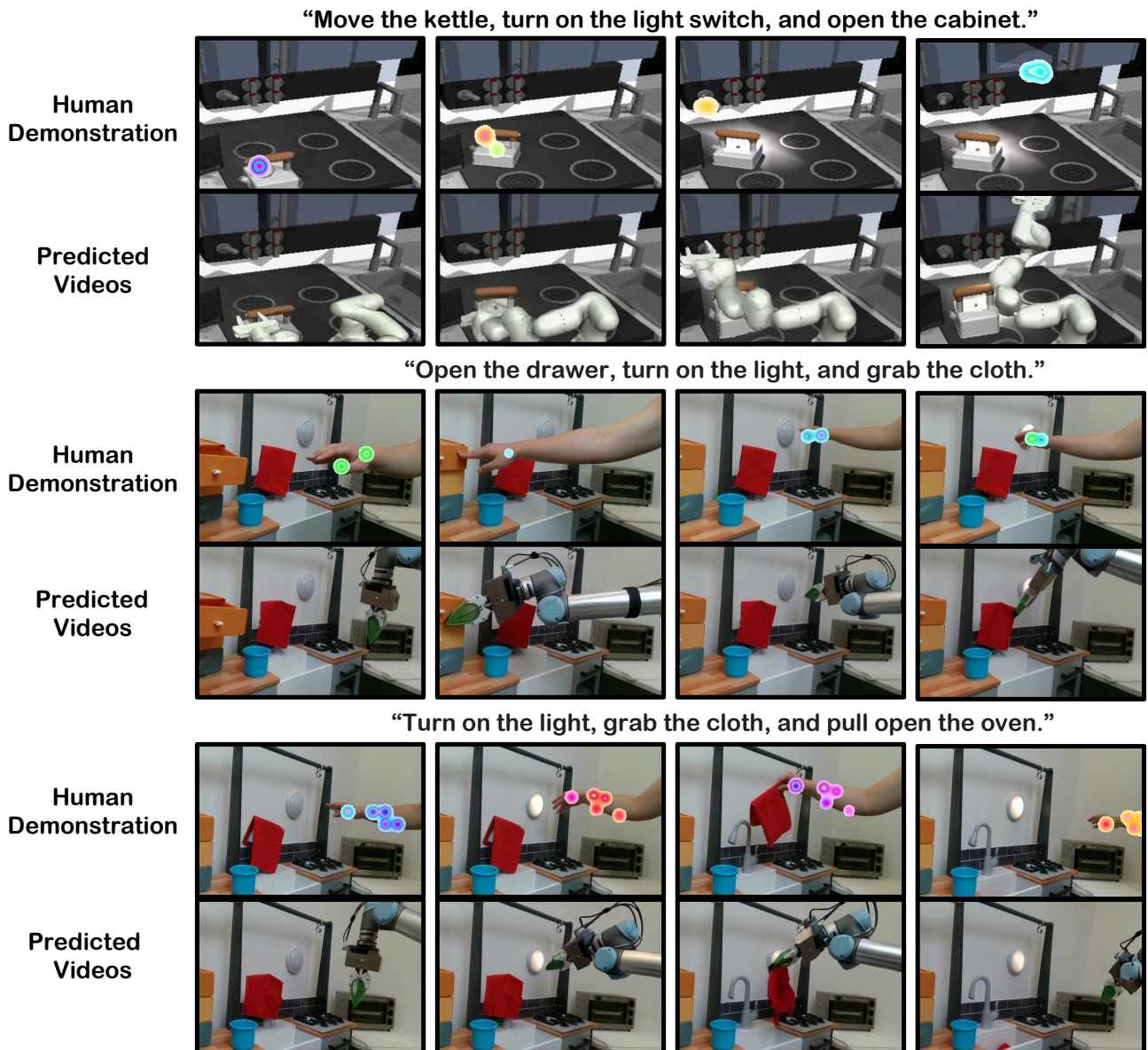


Fig. 4: Trajectory conditioned cross-embodiment skill transfer. Top two rows show simulation results where human demonstrations are abstracted as spherical trajectories (first row) to guide robotic arm motion generation (second row). Bottom four rows demonstrate real-world transfer from human hand demonstrations to robotic arm execution for complex multi-step tasks.

Specifically, our method excels in both easy and hard tasks, reaching 81.8% and 38.0% SR respectively, while maintaining strong performance in the very hard category (30.0%). In contrast, TinyVLA and SmolVLA perform competitively on medium tasks but drop significantly on hard tasks. Diffusion Policy shows limited generalization across all difficulty levels, with only 10.5% overall SR. These results highlight that trajectory conditioned video generation not only enables execution but also scales effectively to more challenging scenarios, offering robust performance across varying task complexities. our method achieves competitive SR compared to Diffusion Policy, TinyVLA, and SmolVLA, verifying that trajectory conditioned video generation provides robot execution.

2) *Real-Robot Experiments:* We deploy our TrajSkill on a Franka Panda in a kitchen environment. As illustrated in Fig. 5, a banana is placed at varying distances (20 cm, 30 cm, 40 cm), and human hand videos provide sparse trajectories to guide motion generation. TrajSkill then transforms the generated video policy into executable robot actions. As shown in Table III, TrajSkill achieves the highest success rates in both pick and place tasks, with 90.9% and 81.8% respectively, which shows that TrajSkill significantly outperforms DP and OCTO in both picking and placing success rates. These results confirm that TrajSkill not only generalizes well in simulation but also transfers effectively to real-world robot execution, achieving robust performance in challenging manipulation scenarios.

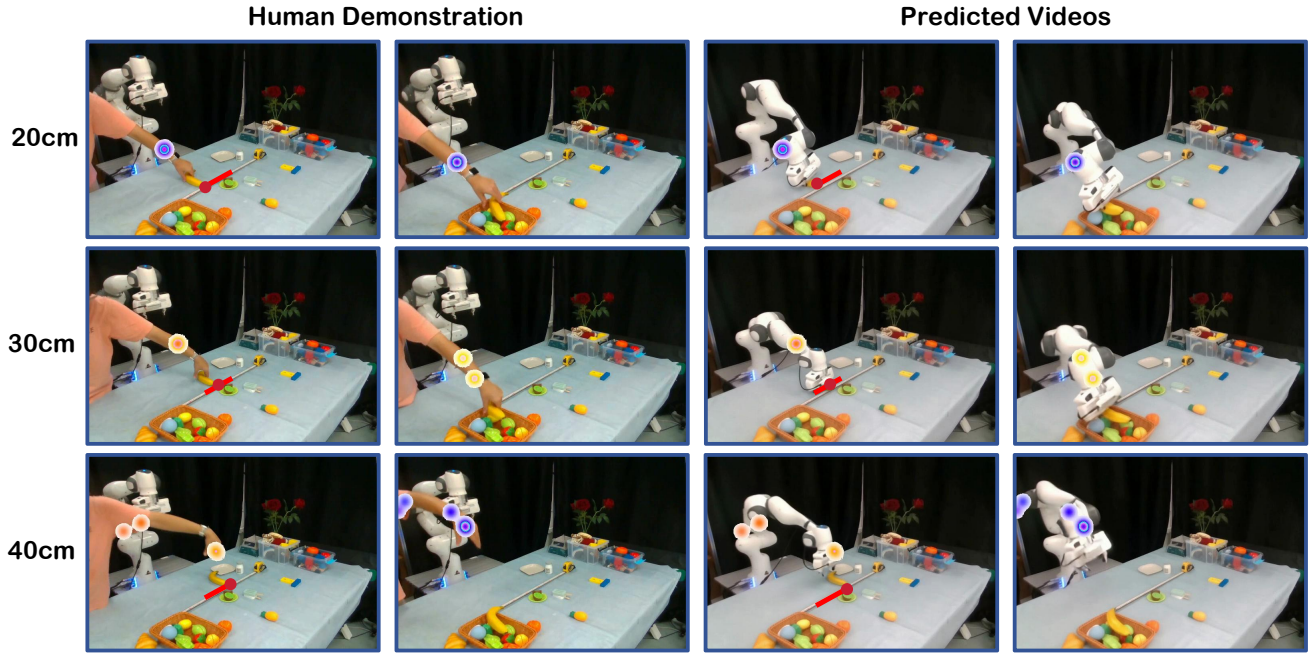


Fig. 5: Human-Controlled Robot Video Prediction for Pick and Place Tasks. Human demonstrations (left) control robot arm movements in predicted videos (right) at three different banana positions: 20cm, 30cm, and 40cm from the basket.

## V. CONCLUSIONS

In this work, we introduce TrajSkill, a trajectory conditioned framework for cross-embodiment skill transfer. Our key idea is to leverage sparse optical flow trajectories extracted from human demonstrations as an embodiment-invariant representation of motion intent. Through trajectory conditioned robot execution, TrajSkill enables direct mapping from human demonstrations to robotic execution, effectively bridging the embodiment gap. Extensive experiments validate the effectiveness of the proposed framework. This work suggests a promising direction for scalable robot learning from unstructured human video demonstrations.

Future work involves extending trajectory-based conditioning to more complex long-horizon tasks, incorporating language grounding for more detailed task specifications, and applying the framework to a variety of robot morphologies in open-world environments.

## ACKNOWLEDGMENTS

This work was supported in part by the National Key Research and Development Project (2024YFC3015503), the National Natural Science Foundation of China (62376222), and the Natural Science Basic Research Program of Shaanxi (2025JC-TBZC-07). This work was supported by NOKOV Motion Capture. We sincerely thank NOKOV Motion Capture for their assistance in our experiments.

## REFERENCES

[1] Y. Liu, A. Gupta, P. Abbeel, and S. Levine, "Imitation from observation: Learning to imitate behaviors from raw video via context translation," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 1118–1125.

[2] L. Shao, T. Migimatsu, Q. Zhang, K. Yang, and J. Bohg, "Concept2robot: Learning manipulation concepts from instructions and human demonstrations," *The International Journal of Robotics Research*, vol. 40, no. 12-14, pp. 1419–1434, 2021.

[3] T. Yu, P. Abbeel, S. Levine, and C. Finn, "One-shot hierarchical imitation learning of compound visuomotor tasks," *arXiv preprint arXiv:1810.11043*, 2018.

[4] A. Majumdar, K. Yadav, S. Arnaud, J. Ma, C. Chen, S. Silwal, A. Jain, V.-P. Berges, T. Wu, J. Vakil, *et al.*, "Where are we in the search for an artificial visual cortex for embodied intelligence?" *Advances in Neural Information Processing Systems*, vol. 36, pp. 655–677, 2023.

[5] F. Bao, C. Xiang, G. Yue, G. He, H. Zhu, K. Zheng, M. Zhao, S. Liu, Y. Wang, and J. Zhu, "Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models," *arXiv preprint arXiv:2405.04233*, 2024.

[6] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, *et al.*, "Video generation models as world simulators," *OpenAI Blog*, vol. 1, no. 8, p. 1, 2024.

[7] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng, *et al.*, "Cogvideox: Text-to-video diffusion models with an expert transformer," *arXiv preprint arXiv:2408.06072*, 2024.

[8] H. Xiong, Q. Li, Y.-C. Chen, H. Bharadhwaj, S. Sinha, and A. Garg, "Learning by watching: Physical imitation of manipulation skills from human videos," in *2021 IEEE/RSJ international conference on intelligent robots and systems (iros)*. IEEE, 2021, pp. 7827–7834.

[9] V. Jain, M. Attarian, N. J. Joshi, A. Wahid, D. Driess, Q. Vuong, P. R. Sanketi, P. Sermanet, S. Welker, C. Chan, *et al.*, "Vid2robot: End-to-end video-conditioned policy learning with cross-attention transformers," *arXiv preprint arXiv:2403.12943*, 2024.

[10] S. Xie, H. Cao, Z. Weng, Z. Xing, H. Chen, S. Shen, J. Leng, Z. Wu, and Y.-G. Jiang, "Human2robot: Learning robot actions from paired human-robot videos," *arXiv preprint arXiv:2502.16587*, 2025.

[11] J. Ren, P. Sundaresan, D. Sadigh, S. Choudhury, and J. Bohg, "Motion tracks: A unified representation for human-robot transfer in few-shot imitation learning," *arXiv preprint arXiv:2501.06994*, 2025.

[12] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, *et al.*, "Imagen video: High definition video generation with diffusion models," *arXiv preprint arXiv:2210.02303*, 2022.

[13] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet,

- “Video diffusion models,” *Advances in neural information processing systems*, vol. 35, pp. 8633–8646, 2022.
- [14] H. Chen, M. Xia, Y. He, Y. Zhang, X. Cun, S. Yang, J. Xing, Y. Liu, Q. Chen, X. Wang, *et al.*, “Videocrafter1: Open diffusion models for high-quality video generation,” *arXiv preprint arXiv:2310.19512*, 2023.
- [15] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voletii, A. Letts, *et al.*, “Stable video diffusion: Scaling latent video diffusion models to large datasets,” *arXiv preprint arXiv:2311.15127*, 2023.
- [16] A. Polyak, A. Zohar, A. Brown, A. Tjandra, A. Sinha, A. Lee, A. Vyas, B. Shi, C.-Y. Ma, C.-Y. Chuang, *et al.*, “Movie gen: A cast of media foundation models,” *arXiv preprint arXiv:2410.13720*, 2024.
- [17] W. W. Liu, J. Keppo, and M. Z. Shou, “Motiondirector: Motion customization of text-to-video diffusion models.”
- [18] H. Jeong, G. Y. Park, and J. C. Ye, “Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9212–9221.
- [19] X. Wang, H. Yuan, S. Zhang, D. Chen, J. Wang, Y. Zhang, Y. Shen, D. Zhao, and J. Zhou, “Videocomposer: Compositional video synthesis with motion controllability,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 7594–7611, 2023.
- [20] Z. Dai, Z. Zhang, Y. Yao, B. Qiu, S. Zhu, L. Qin, and W. Wang, “Fine-grained open domain image animation with motion guidance,” *CoRR*, 2023.
- [21] W. Wu, Z. Li, Y. Gu, R. Zhao, Y. He, D. J. Zhang, M. Z. Shou, Y. Li, T. Gao, and D. Zhang, “Draganything: Motion control for anything using entity representation,” in *European Conference on Computer Vision*. Springer, 2024, pp. 331–348.
- [22] S. Yin, C. Wu, J. Liang, J. Shi, H. Li, G. Ming, and N. Duan, “Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory,” *arXiv preprint arXiv:2308.08089*, 2023.
- [23] Z. Wang, Z. Yuan, X. Wang, Y. Li, T. Chen, M. Xia, P. Luo, and Y. Shan, “Motionctrl: A unified and flexible motion controller for video generation,” in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11.
- [24] M. Yang, Y. Du, K. Ghasemipour, J. Tompson, D. Schuurmans, and P. Abbeel, “Learning interactive real-world simulators,” *arXiv preprint arXiv:2310.06114*, vol. 1, no. 2, p. 6, 2023.
- [25] Y. Du, S. Yang, B. Dai, H. Dai, O. Nachum, J. Tenenbaum, D. Schuurmans, and P. Abbeel, “Learning universal policies via text-guided video generation,” *Advances in neural information processing systems*, vol. 36, pp. 9156–9172, 2023.
- [26] K. Black, M. Nakamoto, P. Atreya, H. Walke, C. Finn, A. Kumar, and S. Levine, “Zero-shot robotic manipulation with pretrained image-editing diffusion models,” *arXiv preprint arXiv:2310.10639*, 2023.
- [27] P.-C. Ko, J. Mao, Y. Du, S.-H. Sun, and J. B. Tenenbaum, “Learning to act from actionless videos through dense correspondences,” *arXiv preprint arXiv:2310.08576*, 2023.
- [28] X. Gu, C. Wen, W. Ye, J. Song, and Y. Gao, “Seer: Language instructed video prediction with latent diffusion models,” *arXiv preprint arXiv:2303.14897*, 2023.
- [29] B. Wang, N. Sridhar, C. Feng, M. Van der Merwe, A. Fishman, N. Fazeli, and J. J. Park, “This&that: Language-gesture controlled video generation for robot planning,” *arXiv preprint arXiv:2407.05530*, 2024.
- [30] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, “R3m: A universal visual representation for robot manipulation,” *arXiv preprint arXiv:2203.12601*, 2022.
- [31] S. Liu, S. Tripathi, S. Majumdar, and X. Wang, “Joint hand motion and interaction hotspots prediction from egocentric videos,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3282–3292.
- [32] M. Goyal, S. Modi, R. Goyal, and S. Gupta, “Human hands as probes for interactive object understanding,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3293–3303.
- [33] C. Wang, L. Fan, J. Sun, R. Zhang, L. Fei-Fei, D. Xu, Y. Zhu, and A. Anandkumar, “Mimicplay: Long-horizon imitation learning by watching human play,” *arXiv preprint arXiv:2302.12422*, 2023.
- [34] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik, “Reconstructing hands in 3d with transformers,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9826–9836.
- [35] Z. Teed and J. Deng, “Raft: Recurrent all-pairs field transforms for optical flow,” in *European conference on computer vision*. Springer, 2020, pp. 402–419.
- [36] R. McLean, E. Chatzaroulas, L. McCutcheon, F. Röder, T. Yu, Z. He, K. Zentner, R. Julian, J. Terry, I. Woungang, *et al.*, “Meta-world+: An improved, standardized, rl benchmark,” *arXiv preprint arXiv:2505.11289*, 2025.
- [37] Y. Seo, D. Hafner, H. Liu, F. Liu, S. James, K. Lee, and P. Abbeel, “Masked world models for visual control,” in *Conference on Robot Learning*. PMLR, 2023, pp. 1332–1344.
- [38] H. Song, D. Qu, Y. Yao, Q. Chen, Q. Lv, Y. Tang, M. Shi, G. Ren, M. Yao, B. Zhao, *et al.*, “Hume: Introducing system-2 thinking in visual-language-action model,” *arXiv preprint arXiv:2505.21432*, 2025.
- [39] M. Xu, Z. Xu, C. Chi, M. Veloso, and S. Song, “Xskill: Cross embodiment skill discovery,” in *Conference on robot learning*. PMLR, 2023, pp. 3536–3555.
- [40] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, “Diffusion policy: Visuomotor policy learning via action diffusion,” *The International Journal of Robotics Research*, p. 02783649241273668, 2023.
- [41] J. Wen, Y. Zhu, J. Li, M. Zhu, Z. Tang, K. Wu, Z. Xu, N. Liu, R. Cheng, C. Shen, *et al.*, “Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation,” *IEEE Robotics and Automation Letters*, 2025.
- [42] M. Shukor, D. Aubakirova, F. Capuano, P. Kooijmans, S. Palma, A. Zouitine, M. Aractingi, C. Pascal, M. Russi, A. Marafioti, *et al.*, “Smolvla: A vision-language-action model for affordable and efficient robotics,” *arXiv preprint arXiv:2506.01844*, 2025.
- [43] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, *et al.*, “Octo: An open-source generalist robot policy,” *arXiv preprint arXiv:2405.12213*, 2024.
- [44] T. Unterthiner, S. Van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, “Towards accurate generative models of video: A new metric & challenges,” *arXiv preprint arXiv:1812.01717*, 2018.
- [45] D. Qu, H. Song, Q. Chen, Y. Yao, X. Ye, Y. Ding, Z. Wang, J. Gu, B. Zhao, D. Wang, *et al.*, “Spatialvla: Exploring spatial representations for visual-language-action model,” *arXiv preprint arXiv:2501.15830*, 2025.